DSC520-FinalProject

SumbarajuAditya

2/25/2021

# library

library(knitr)  
library(ggplot2)  
library(gvlma)  
library(pander)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

#library(reshape2)   
#library(ISLR)   
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

#library(leaps)  
library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.1

#library(bestglm)   
#library(pROC)  
#library(broom)  
#library(mapproj)  
#library(boot)  
#library(tree)  
#library(randomForest)   
#library(rpart)

Data importing and cleaning steps

filename <- 'C:/BU/DSC520/assignment\_repo/dsc520/data/final\_project/diabetic\_data\_v1.csv'  
diabetic\_data <- read.table(filename, sep = ",", header = T, na.strings = "?")  
#head(data)

For getting the readmission prediction we dont need many columns hence we should be selecting only relevant values to make it simpler.Also we need to remove those columns which mainly has NA values.

diabetic\_data$admission\_type\_id <- as.factor(diabetic\_data$admission\_type\_id)  
diabetic\_data$discharge\_disposition\_id <- as.factor(diabetic\_data$discharge\_disposition\_id)  
diabetic\_data$admission\_source\_id <- as.factor(diabetic\_data$admission\_source\_id)

Cleaning Missing Values Below is the logic to replace bad data “?” and “unknown/invalid values” with NA. I am passing respective columns from diabetic\_data\_clean “diabetic\_data” using ncol and trying to identify the bad data by passing the columns in loop and there by replacing the data with NA.

count <- 0  
for(i in 1:ncol(diabetic\_data))  
 {  
 if(is.factor(diabetic\_data[,i]))  
 {  
 for(j in 1:nrow(diabetic\_data))  
 {  
 if(diabetic\_data[j,i]== "?" | diabetic\_data[j,i]== "Unknown/Invalid" )  
 {  
 count <- count + 1  
 diabetic\_data[j,i] <- NA #replace "?" and "Unknown/Invalid" values with NA  
 }  
 }  
 if(count > 0)  
 {  
 print(c(colnames(diabetic\_data)[i],count))  
 }  
 }  
 count <- 0  
 }

Below are the variables removed: Payer code, weight and Medical Specialty are not included since they have a large number of missing values.

Minor variability variables are excluded. That includes following variables acetohexamide, glimepiride.pioglitazone, metformin.rosiglitazone, metformin.pioglitazone, chlorpropamide, acetohexamide, tolbutamide, acarbose, miglitor, troglitazone, tolazamide, examide, citoglipton, glyburide.metformin, glipizide.metformin, and glimepiride.pioglitazone.

diabetic\_data\_clean <- diabetic\_data  
diabetic\_data\_clean$X <- NULL  
diabetic\_data\_clean$weight <- NULL  
diabetic\_data\_clean$payer\_code <- NULL  
diabetic\_data\_clean$admission\_type\_id <- NULL  
diabetic\_data\_clean$discharge\_disposition\_id <- NULL  
diabetic\_data\_clean$admission\_source\_id <- NULL   
diabetic\_data\_clean$medical\_specialty <- NULL  
diabetic\_data\_clean$encounter\_id <- NULL  
diabetic\_data\_clean$diag\_1 <- NULL  
diabetic\_data\_clean$diag\_2 <- NULL  
diabetic\_data\_clean$diag\_3 <- NULL  
diabetic\_data\_clean$examide <- NULL  
diabetic\_data\_clean$citoglipton <- NULL  
diabetic\_data\_clean$acetohexamide <- NULL  
diabetic\_data\_clean$repaglinide <- NULL  
diabetic\_data\_clean$nateglinide <- NULL  
diabetic\_data\_clean$metformin.pioglitazone <- NULL   
diabetic\_data\_clean$metformin.rosiglitazone <- NULL  
diabetic\_data\_clean$chlorpropamide <- NULL  
diabetic\_data\_clean$acetohexamide <- NULL  
diabetic\_data\_clean$miglitol <- NULL   
diabetic\_data\_clean$tolbutamide <- NULL  
diabetic\_data\_clean$acarbose <- NULL  
diabetic\_data\_clean$miglitor <- NULL  
diabetic\_data\_clean$troglitazone <- NULL  
diabetic\_data\_clean$tolazamide <- NULL  
diabetic\_data\_clean$examide <- NULL  
diabetic\_data\_clean$citoglipton <- NULL  
diabetic\_data\_clean$glyburide.metformin <- NULL  
diabetic\_data\_clean$glipizide.metformin <- NULL  
diabetic\_data\_clean$glimepiride.pioglitazone <- NULL  
diabetic\_data\_clean <- na.omit(diabetic\_data\_clean)  
#write.csv(diabetic\_data\_clean, file = "C:/BU/DSC520/assignment\_repo/dsc520/data/final\_project/diabetic\_data\_clean2.csv")  
dim(diabetic\_data\_clean)

## [1] 98144 30

Below is the Summary & Structure of the data;

str(diabetic\_data\_clean)

## 'data.frame': 98144 obs. of 30 variables:  
## $ patient\_nbr : int 48330783 63555939 42519267 89869032 82637451 77391171 85504905 77586282 84259809 49726791 ...  
## $ race : chr "Caucasian" "Caucasian" "Caucasian" "AfricanAmerican" ...  
## $ gender : chr "Female" "Female" "Male" "Female" ...  
## $ age\_mod : chr "80+" "80+" "20-59" "20-59" ...  
## $ time\_in\_hospital : int 13 12 1 9 3 7 7 10 4 1 ...  
## $ num\_lab\_procedures : int 68 33 51 47 31 62 60 55 70 49 ...  
## $ num\_procedures : int 2 3 0 2 6 0 0 1 1 5 ...  
## $ num\_medications : int 28 18 8 17 16 11 15 31 21 2 ...  
## $ number\_outpatient : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ number\_emergency : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ number\_inpatient : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ number\_diagnoses : int 8 8 5 9 9 7 8 8 7 8 ...  
## $ max\_glu\_serum : chr "None" "None" "None" "None" ...  
## $ A1Cresult : chr "None" "None" "None" "None" ...  
## $ metformin : chr "No" "No" "No" "No" ...  
## $ glimepiride : chr "No" "No" "No" "No" ...  
## $ glipizide : chr "Steady" "No" "Steady" "No" ...  
## $ glyburide : chr "No" "No" "No" "No" ...  
## $ pioglitazone : chr "No" "No" "No" "No" ...  
## $ rosiglitazone : chr "No" "Steady" "No" "No" ...  
## $ insulin : chr "Steady" "Steady" "Steady" "Steady" ...  
## $ change : chr "Ch" "Ch" "Ch" "No" ...  
## $ diabetesMed : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ disch\_disp\_modified: chr "Discharged to home" "Discharged/Transferred to SNF" "Discharged to home" "Discharged to home" ...  
## $ adm\_src\_mod : chr "Other" "Other" "Emergency Room" "Emergency Room" ...  
## $ adm\_typ\_mod : chr "Urgent" "Elective" "Emergency" "Emergency" ...  
## $ diag1\_mod : chr "Other" "434" "Other" "Other" ...  
## $ diag2\_mod : chr "427" "Other" "Other" "403" ...  
## $ diag3\_mod : chr "Other" "Other" "250" "Other" ...  
## $ readmitted : chr "NO" "NO" "NO" ">30" ...  
## - attr(\*, "na.action")= 'omit' Named int [1:3622] 17 19 20 54 65 66 87 99 111 127 ...  
## ..- attr(\*, "names")= chr [1:3622] "17" "19" "20" "54" ...

#head (diabetic\_data\_clean)  
# Raw diabetic\_data\_clean:  
dim(diabetic\_data)

## [1] 101766 56

# Clean diabetic\_data\_clean  
dim(diabetic\_data\_clean)

## [1] 98144 30

#write.csv(diabetic\_data\_clean, file = "diabetic\_data\_clean1.csv")

#diabetic\_data\_clean <- subset(diabetic\_data\_clean, select = -c(encounter\_id)) # delete ID variables  
names(diabetic\_data\_clean)[length(diabetic\_data\_clean)] <- "y"   
diabetic\_data\_clean$y <- as.character(diabetic\_data\_clean$y)  
diabetic\_data\_clean$y[which(diabetic\_data\_clean$y == ">30" | diabetic\_data\_clean$y == "NO")] <- "0"  
diabetic\_data\_clean$y[which(diabetic\_data\_clean$y == "<30")] <- "1"  
diabetic\_data\_clean$y <- as.factor(diabetic\_data\_clean$y)  
  
#rename variables  
names(diabetic\_data\_clean) <- c("pers", "race", "gend", "agem", "hosp", "labp", "proc", "nmed", "nout", "emer", "ninp", "diag", "mglu", "A1Cr", "metf", "glim", "glip", "glyb", "piog", "rosi", "insu", "chge", "diab", "disc", "adms", "admt", "mod1", "mod2", "mod3", "y")  
  
str(diabetic\_data\_clean)

## 'data.frame': 98144 obs. of 30 variables:  
## $ pers: int 48330783 63555939 42519267 89869032 82637451 77391171 85504905 77586282 84259809 49726791 ...  
## $ race: chr "Caucasian" "Caucasian" "Caucasian" "AfricanAmerican" ...  
## $ gend: chr "Female" "Female" "Male" "Female" ...  
## $ agem: chr "80+" "80+" "20-59" "20-59" ...  
## $ hosp: int 13 12 1 9 3 7 7 10 4 1 ...  
## $ labp: int 68 33 51 47 31 62 60 55 70 49 ...  
## $ proc: int 2 3 0 2 6 0 0 1 1 5 ...  
## $ nmed: int 28 18 8 17 16 11 15 31 21 2 ...  
## $ nout: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ emer: int 0 0 0 0 0 0 1 0 0 0 ...  
## $ ninp: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ diag: int 8 8 5 9 9 7 8 8 7 8 ...  
## $ mglu: chr "None" "None" "None" "None" ...  
## $ A1Cr: chr "None" "None" "None" "None" ...  
## $ metf: chr "No" "No" "No" "No" ...  
## $ glim: chr "No" "No" "No" "No" ...  
## $ glip: chr "Steady" "No" "Steady" "No" ...  
## $ glyb: chr "No" "No" "No" "No" ...  
## $ piog: chr "No" "No" "No" "No" ...  
## $ rosi: chr "No" "Steady" "No" "No" ...  
## $ insu: chr "Steady" "Steady" "Steady" "Steady" ...  
## $ chge: chr "Ch" "Ch" "Ch" "No" ...  
## $ diab: chr "Yes" "Yes" "Yes" "Yes" ...  
## $ disc: chr "Discharged to home" "Discharged/Transferred to SNF" "Discharged to home" "Discharged to home" ...  
## $ adms: chr "Other" "Other" "Emergency Room" "Emergency Room" ...  
## $ admt: chr "Urgent" "Elective" "Emergency" "Emergency" ...  
## $ mod1: chr "Other" "434" "Other" "Other" ...  
## $ mod2: chr "427" "Other" "Other" "403" ...  
## $ mod3: chr "Other" "Other" "250" "Other" ...  
## $ y : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 1 1 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:3622] 17 19 20 54 65 66 87 99 111 127 ...  
## ..- attr(\*, "names")= chr [1:3622] "17" "19" "20" "54" ...

# Data partition  
set.seed(1121)  
pos <- sample(1:nrow(diabetic\_data\_clean), round(nrow(diabetic\_data\_clean)\*0.2))  
testing <- diabetic\_data\_clean[pos, -1]   
training <- diabetic\_data\_clean[-pos, -1]   
  
# XX and y   
data <- training  
y <- data$y  
response <- "readmitted"   
XX <- model.matrix(y ~., data)[, -1]  
Y <- y  
glimpse(data)

## Rows: 78,515  
## Columns: 29  
## $ race <chr> "Caucasian", "Caucasian", "AfricanAmerican", "AfricanAmerican"...  
## $ gend <chr> "Female", "Female", "Female", "Male", "Female", "Male", "Male"...  
## $ agem <chr> "80+", "80+", "20-59", "60-79", "20-59", "80+", "60-79", "60-7...  
## $ hosp <int> 13, 12, 9, 7, 7, 10, 4, 5, 2, 12, 4, 3, 3, 2, 3, 1, 6, 2, 10, ...  
## $ labp <int> 68, 33, 47, 62, 60, 55, 70, 73, 11, 75, 45, 29, 59, 47, 19, 33...  
## $ proc <int> 2, 3, 2, 0, 0, 1, 1, 0, 5, 5, 4, 0, 0, 0, 4, 0, 3, 2, 0, 0, 0,...  
## $ nmed <int> 28, 18, 17, 11, 15, 31, 21, 12, 13, 13, 17, 11, 18, 12, 18, 7,...  
## $ nout <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ emer <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ ninp <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ diag <int> 8, 8, 9, 7, 8, 8, 7, 8, 6, 9, 8, 3, 9, 8, 6, 3, 7, 3, 6, 8, 9,...  
## $ mglu <chr> "None", "None", "None", "None", "None", "None", "None", "None"...  
## $ A1Cr <chr> "None", "None", "None", "None", "None", "None", "None", "None"...  
## $ metf <chr> "No", "No", "No", "No", "Steady", "No", "Steady", "No", "No", ...  
## $ glim <chr> "No", "No", "No", "No", "No", "No", "Steady", "No", "No", "No"...  
## $ glip <chr> "Steady", "No", "No", "No", "No", "No", "No", "No", "Steady", ...  
## $ glyb <chr> "No", "No", "No", "Up", "No", "No", "No", "Steady", "No", "No"...  
## $ piog <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "N...  
## $ rosi <chr> "No", "Steady", "No", "No", "No", "No", "No", "No", "No", "No"...  
## $ insu <chr> "Steady", "Steady", "Steady", "Steady", "Down", "Steady", "Ste...  
## $ chge <chr> "Ch", "Ch", "No", "Ch", "Ch", "No", "Ch", "No", "No", "Ch", "C...  
## $ diab <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",...  
## $ disc <chr> "Discharged to home", "Discharged/Transferred to SNF", "Discha...  
## $ adms <chr> "Other", "Other", "Emergency Room", "Other", "Emergency Room",...  
## $ admt <chr> "Urgent", "Elective", "Emergency", "Urgent", "Emergency", "Eme...  
## $ mod1 <chr> "Other", "434", "Other", "Other", "428", "428", "414", "428", ...  
## $ mod2 <chr> "427", "Other", "403", "Other", "Other", "411", "411", "Other"...  
## $ mod3 <chr> "Other", "Other", "Other", "Other", "250.6", "427", "V45", "25...  
## $ y <fct> 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

glimpse(testing)

## Rows: 19,629  
## Columns: 29  
## $ race <chr> "AfricanAmerican", "Caucasian", "AfricanAmerican", "Caucasian"...  
## $ gend <chr> "Male", "Male", "Female", "Female", "Male", "Male", "Female", ...  
## $ agem <chr> "20-59", "20-59", "60-79", "80+", "60-79", "60-79", "20-59", "...  
## $ hosp <int> 1, 1, 1, 1, 3, 3, 5, 11, 2, 7, 3, 8, 2, 2, 2, 2, 9, 8, 2, 3, 1...  
## $ labp <int> 54, 49, 38, 45, 44, 41, 42, 45, 42, 81, 18, 33, 57, 39, 50, 42...  
## $ proc <int> 0, 2, 2, 0, 0, 0, 0, 0, 0, 4, 0, 3, 0, 0, 0, 4, 0, 3, 0, 5, 4,...  
## $ nmed <int> 10, 19, 28, 6, 16, 15, 22, 29, 6, 19, 13, 4, 12, 3, 23, 24, 20...  
## $ nout <int> 0, 0, 0, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...  
## $ emer <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 1, 0,...  
## $ ninp <int> 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 5, 0, 4, 1, 2, 0, 1,...  
## $ diag <int> 3, 9, 3, 4, 9, 9, 9, 7, 4, 9, 9, 9, 7, 6, 9, 7, 9, 9, 9, 9, 5,...  
## $ mglu <chr> "None", "None", "None", "None", "None", ">300", "None", "None"...  
## $ A1Cr <chr> ">8", ">8", "None", "None", "None", "None", "None", "None", "N...  
## $ metf <chr> "No", "Steady", "Steady", "No", "No", "No", "No", "No", "No", ...  
## $ glim <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "N...  
## $ glip <chr> "No", "No", "No", "Steady", "No", "No", "No", "Steady", "No", ...  
## $ glyb <chr> "No", "No", "Steady", "No", "No", "Steady", "No", "No", "No", ...  
## $ piog <chr> "No", "Steady", "No", "No", "No", "No", "No", "No", "No", "No"...  
## $ rosi <chr> "No", "No", "No", "Steady", "No", "No", "No", "No", "No", "No"...  
## $ insu <chr> "Steady", "Steady", "No", "No", "No", "No", "Up", "Steady", "N...  
## $ chge <chr> "No", "Ch", "Ch", "Ch", "No", "No", "Ch", "Ch", "No", "Ch", "N...  
## $ diab <chr> "Yes", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", "...  
## $ disc <chr> "Discharged to home", "Discharged to home", "Discharged to hom...  
## $ adms <chr> "Emergency Room", "Emergency Room", "Physician Referral", "Eme...  
## $ admt <chr> "Emergency", "Emergency", "Elective", "Emergency", "Emergency"...  
## $ mod1 <chr> "Other", "414", "Other", "780", "428", "434", "250.6", "Other"...  
## $ mod2 <chr> "Other", "411", "250", "414", "414", "428", "585", "Other", "4...  
## $ mod3 <chr> "Other", "Other", "401", "250", "250", "585", "403", "428", "2...  
## $ y <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

#Patient Demographics

Readmission\_indicator <- prop.table(table(data$y))   
yprop2 <- with(data, table(y, y))  
yprop3 <- prop.table(yprop2, margin = 1)   
df.y <- data.frame(round(rbind(Readmission\_indicator, t(yprop3)), 3)\*100)  
names(df.y) <- c("Never or > 30 days","< 30 days")  
head (df.y,n=1)

## Never or > 30 days < 30 days  
## Readmission\_indicator 88.7 11.3

About 11.3% of observations have a record of readmission within 30 days. 88.7 % of observations have a record of readmission >30 days.

# subset data which excludes duplicate data points  
# demographics  
Total <- prop.table(table(data$race))   
crosstab <- with(data, table(race, y))  
ytab2 <- prop.table(crosstab, margin = 2)   
df.race <- data.frame(round(rbind(Total, t(ytab2)), 3))  
#names(df.race)[1] <- "Missing"  
#rownames(df.race) <- c("Total", "y = 0", "y = 1")  
df.race

## AfricanAmerican Asian Caucasian Hispanic Other  
## Total 0.193 0.006 0.766 0.02 0.015  
## 0 0.193 0.006 0.765 0.02 0.016  
## 1 0.192 0.006 0.770 0.02 0.013

Total <- prop.table(table(data$gend))   
crosstab <- with(data, table(gend, y))  
ytab2 <- prop.table(crosstab, margin = 2)   
df.gender <- data.frame(round(rbind(Total, t(ytab2)), 3))  
#rownames(df.gender) <- c("Total","y = 0", "y = 1")  
df.gender

## Female Male Unknown.Invalid  
## Total 0.540 0.460 0  
## 0 0.540 0.460 0  
## 1 0.543 0.457 0

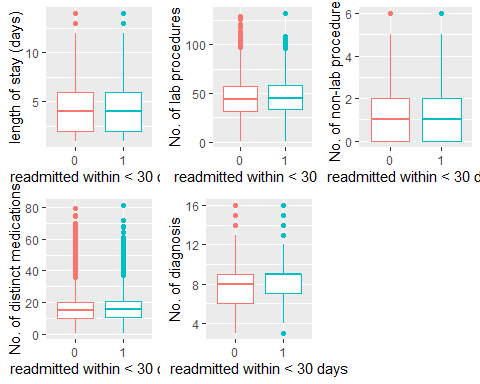
Total <- prop.table(table(data$agem))   
crosstab <- with(data, table(agem, y))  
ytab2 <- prop.table(crosstab, margin = 2)   
df.age <- data.frame(round(rbind(Total, t(ytab2)), 3))  
#rownames(df.age) <- c("Total", "y = 0", "y = 1")  
#colnames(df.age) <- c("0~19", "20~59", "60~79", "80+")  
df.age

## X0.19 X20.59 X60.79 X80.  
## Total 0.006 0.316 0.481 0.198  
## 0 0.006 0.319 0.480 0.196  
## 1 0.003 0.294 0.492 0.211

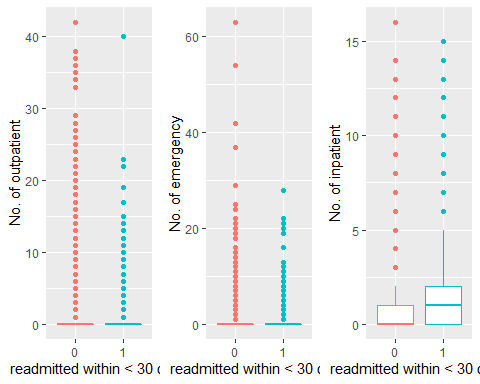
# Patient Medical History

As for labp, proc, and nmed, each shares similar distributions in the meas and spreads under two levels of the readmission status. For hosp and diag, the spread of the length of stay in the hospital is larger and the average number of diagnosis is higher for a patient being readmitted within 30 days.

# names(data)  
# df <- subset(data, select = c(race, gender, y))  
# knitr::kable(df)  
# continuous data  
plot1 <- ggplot(data, aes(x = y, y = hosp, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "length of stay (days)", x = "readmitted within < 30 days")   
plot2 <- ggplot(data, aes(x = y, y = labp, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "No. of lab procedures", x = "readmitted within < 30 days")   
plot3 <- ggplot(data, aes(x = y, y = proc, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "No. of non-lab procedures", x = "readmitted within < 30 days")   
plot4 <- ggplot(data, aes(x = y, y = nmed, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "No. of distinct medications", x = "readmitted within < 30 days")   
plot8 <- ggplot(data, aes(x = y, y = diag, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "No. of diagnosis", x = "readmitted within < 30 days")   
grid.arrange(plot1, plot2, plot3, plot4, plot8, nrow = 2)



plot5 <- ggplot(data, aes(x = y, y = nout, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "No. of outpatient", x = "readmitted within < 30 days")   
plot6 <- ggplot(data, aes(x = y, y = emer, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "No. of emergency", x = "readmitted within < 30 days")   
plot7 <- ggplot(data, aes(x = y, y = ninp, color = y)) + geom\_boxplot() + theme(legend.position = "none") +  
 labs(y = "No. of inpatient", x = "readmitted within < 30 days")   
grid.arrange(plot5, plot6, plot7, nrow = 1)



#Clinical Results: > The correlation between mglu and A1Cr is -0.046, which is not significant.

The diagnosis of diabetes is based on the patients’ blood glucose level, i.e. mglu. And the A1c test (A1Cr) is an important measure of how well a person with diabetes is controlling their blood glucose level. Note: The American Diabetes Association recommends a goal of less than 7.0% A1c.

subdata <- subset(data, select = c(mglu, A1Cr))  
subdata <- as.matrix(subdata)  
subdata[which(subdata == "None")] <- 0  
subdata[which(subdata == "Norm")] <- 1  
subdata[which(subdata == ">7")] <- 2  
subdata[which(subdata == ">8")] <- 3  
subdata[which(subdata == ">200")] <- 2  
subdata[which(subdata == ">300")] <- 3  
subdata <- matrix(as.numeric(subdata), ncol = 2)  
subdata <- data.frame(subdata)  
names(subdata) <- c("mglu", "A1Cr")  
cor <- round(cor(subdata), 3)  
summary(cor)

## mglu A1Cr   
## Min. :-0.0450 Min. :-0.0450   
## 1st Qu.: 0.2162 1st Qu.: 0.2162   
## Median : 0.4775 Median : 0.4775   
## Mean : 0.4775 Mean : 0.4775   
## 3rd Qu.: 0.7388 3rd Qu.: 0.7388   
## Max. : 1.0000 Max. : 1.0000

#Medication Details:

Total <- prop.table(table(data$chge))   
crosstab <- with(data, table(chge, y))  
ytab2 <- prop.table(crosstab, margin = 2)   
df <- data.frame(round(rbind(Total, t(ytab2)), 3))  
#rownames(df) <- c("Total","y = 0", "y = 1")  
names(df) <- c("Change diabetes medication", "No change")  
df

## Change diabetes medication No change  
## Total 0.463 0.537  
## 0 0.460 0.540  
## 1 0.489 0.511

# relationship between medication details  
nomed <- sum(with(data, metf == "No" & glim == "No" & glip == "No" & glyb == "No" & piog == "No" & rosi == "No" & insu == "No"))  
subdata <- subset(data, select = c(metf, glim, glip, glyb, piog, rosi, insu))  
subdata <- as.matrix(subdata)  
subdata[which(subdata == "No")] <- 0  
subdata[which(subdata == "Down")] <- 1  
subdata[which(subdata == "Steady")] <- 2  
subdata[which(subdata == "Up")] <- 3  
subdata <- matrix(as.numeric(subdata), ncol = 7)  
subdata <- data.frame(subdata)  
names(subdata) <- c("metf", "glim", "glip", "glyb", "piog", "rosi", "insu")  
data.frame(round(cor(subdata), 3))

## metf glim glip glyb piog rosi insu  
## metf 1.000 0.042 0.077 0.136 0.059 0.094 -0.026  
## glim 0.042 1.000 -0.070 -0.067 0.042 0.037 0.011  
## glip 0.077 -0.070 1.000 -0.105 0.045 0.044 -0.026  
## glyb 0.136 -0.067 -0.105 1.000 0.023 0.035 -0.067  
## piog 0.059 0.042 0.045 0.023 1.000 -0.064 0.004  
## rosi 0.094 0.037 0.044 0.035 -0.064 1.000 0.004  
## insu -0.026 0.011 -0.026 -0.067 0.004 0.004 1.000

#Admission and discharge details

# factor  
Total <- prop.table(table(data$disc))   
crosstab <- with(data, table(disc, y))  
ytab2 <- prop.table(crosstab, margin = 2)   
df.disc <- data.frame(round(rbind(Total, t(ytab2)), 3))  
#rownames(df) <- c("Total","y = 0", "y = 1")  
#names(df.disc) <- c("Discharged\_to\_home","Discharged\_to\_home\_HomeHealthService","Discharged\_Transferred\_to\_SNF","Other")  
  
df.disc

## Discharged.to.home Discharged.to.home.with.Home.Health.Service  
## Total 0.587 0.128  
## 0 0.599 0.126  
## 1 0.495 0.148  
## Discharged.Transferred.to.SNF Other  
## Total 0.139 0.146  
## 0 0.134 0.142  
## 1 0.178 0.180

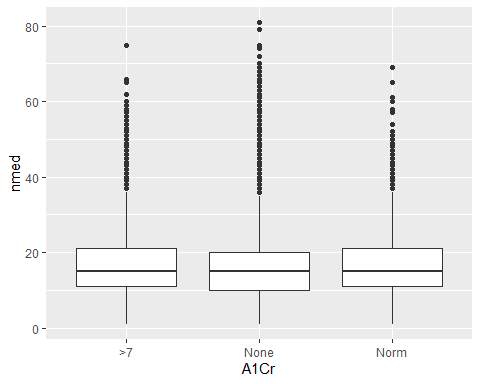
Total <- prop.table(table(data$adms))   
crosstab <- with(data, table(adms, y))  
ytab2 <- prop.table(crosstab, margin = 2)   
df.adms <- data.frame(round(rbind(Total, t(ytab2)), 3))  
#rownames(df) <- c("Total","y = 0", "y = 1")  
#names(df.adms) <- c("Emergency\_Room","Other","Physician\_Referral","Transfer\_from\_Home.Health")  
  
df.adms

## Emergency.Room Other Physician.Referral Transfer.from.Home.Health  
## Total 0.570 0.073 0.289 0.068  
## 0 0.567 0.074 0.290 0.068  
## 1 0.592 0.066 0.278 0.064

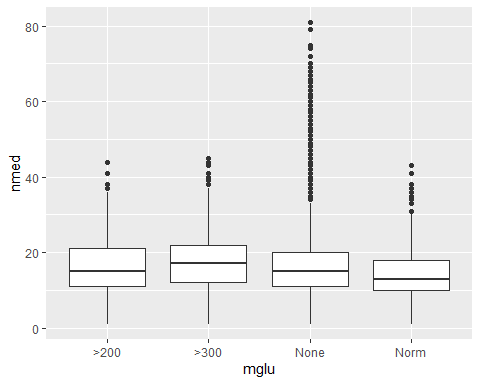
Total <- prop.table(table(data$admt))   
crosstab <- with(data, table(admt, y))  
ytab2 <- prop.table(crosstab, margin = 2)   
df.admt <- data.frame(round(rbind(Total, t(ytab2)), 3))  
#rownames(df) <- c("Total","y = 0", "y = 1")  
df.admt

## Elective Emergency Other Urgent  
## Total 0.185 0.532 0.104 0.178  
## 0 0.187 0.531 0.105 0.178  
## 1 0.175 0.544 0.101 0.181

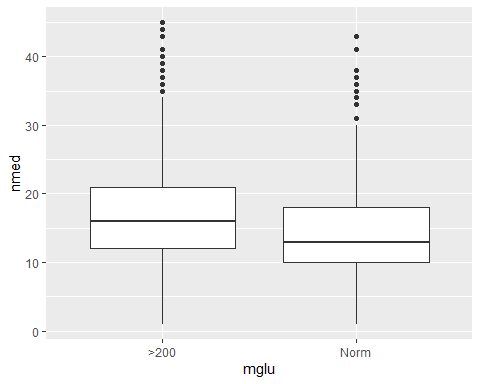
data$A1Cr <- data$A1Cr %>% recode\_factor('>8' = '>7')  
   
ggplot(data = data, aes(x = A1Cr, y = nmed)) + geom\_boxplot()



ggplot(data = data, aes(x = mglu, y = nmed)) + geom\_boxplot()



data$mglu <- data$mglu %>% recode\_factor('>300' = '>200')  
  
data <- data %>% filter(mglu != 'None') %>% drop()  
  
ggplot(data = data, aes(x = mglu, y = nmed)) + geom\_boxplot()

 #car plot

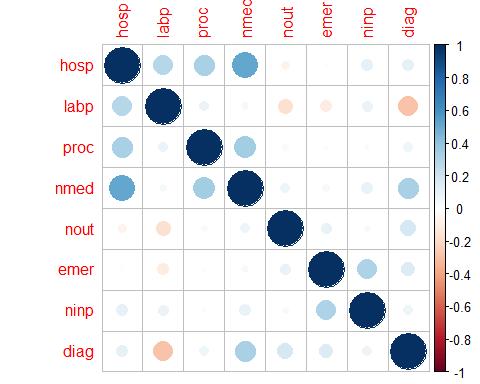
str(data)

## 'data.frame': 4198 obs. of 29 variables:  
## $ race: chr "Caucasian" "Caucasian" "Caucasian" "Caucasian" ...  
## $ gend: chr "Male" "Female" "Male" "Male" ...  
## $ agem: chr "60-79" "80+" "60-79" "80+" ...  
## $ hosp: int 6 11 7 5 6 7 8 7 2 1 ...  
## $ labp: int 55 70 82 47 57 37 35 66 58 59 ...  
## $ proc: int 1 1 2 1 0 2 2 3 0 0 ...  
## $ nmed: int 13 23 21 6 7 9 8 19 9 5 ...  
## $ nout: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ emer: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ninp: int 0 0 0 0 0 1 1 0 0 0 ...  
## $ diag: int 3 5 5 5 5 3 4 4 5 5 ...  
## $ mglu: Factor w/ 3 levels ">200","None",..: 1 3 3 1 1 3 3 1 3 1 ...  
## $ A1Cr: Factor w/ 3 levels ">7","None","Norm": 2 2 2 3 2 2 2 2 2 2 ...  
## $ metf: chr "No" "No" "No" "No" ...  
## $ glim: chr "No" "No" "No" "No" ...  
## $ glip: chr "Up" "No" "No" "No" ...  
## $ glyb: chr "No" "No" "No" "No" ...  
## $ piog: chr "No" "No" "No" "No" ...  
## $ rosi: chr "No" "No" "No" "No" ...  
## $ insu: chr "No" "No" "No" "No" ...  
## $ chge: chr "Ch" "No" "No" "No" ...  
## $ diab: chr "Yes" "No" "No" "No" ...  
## $ disc: chr "Discharged to home" "Discharged/Transferred to SNF" "Discharged to home" "Discharged/Transferred to SNF" ...  
## $ adms: chr "Physician Referral" "Emergency Room" "Emergency Room" "Emergency Room" ...  
## $ admt: chr "Other" "Other" "Other" "Other" ...  
## $ mod1: chr "Other" "Other" "427" "Other" ...  
## $ mod2: chr "250.02" "Other" "Other" "Other" ...  
## $ mod3: chr "276" "276" "Other" "425" ...  
## $ y : Factor w/ 2 levels "0","1": 1 1 1 2 2 1 1 1 1 1 ...

library(corrplot)

## corrplot 0.84 loaded

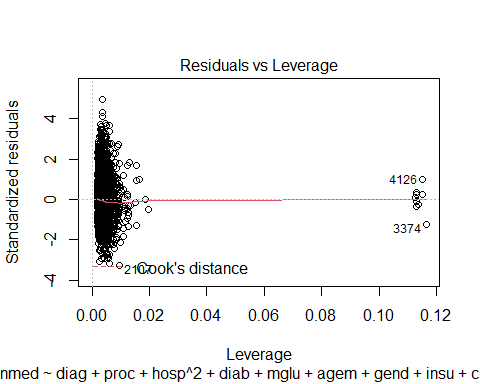
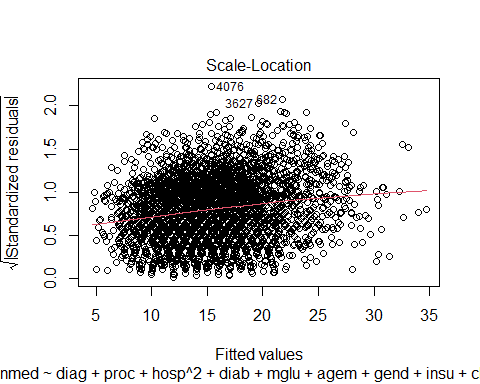
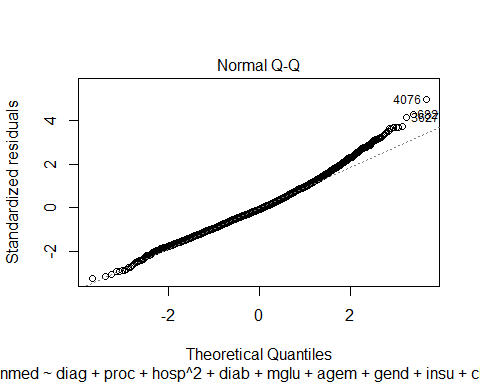
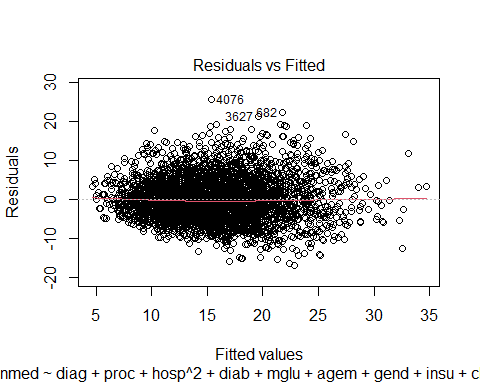
corrplot(cor(data[, unlist(lapply(data, is.numeric))]))

 > There appear to be some positive correlations between number of medications and variables like the time spent in hospital, number of procedures, and diagnoses. Next, I’ll look at the categorical variable age, and see if this conforms to my expectation that diabetic patients trend to be older on average.

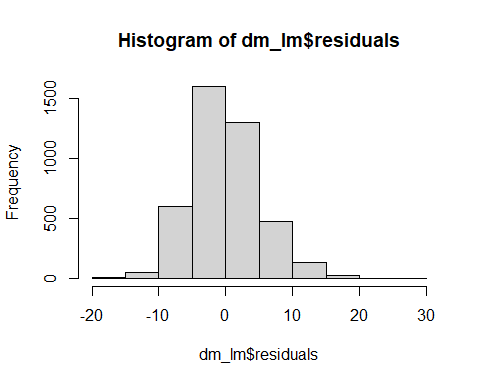
#names(diabetic\_data\_clean) <- c("pers", "race", "gend", "agem", "hosp", "labp", "proc", "nmed", "nout", "emer", "ninp", "diag", "mglu", "A1Cr", "metf", "glim", "glip", "glyb", "piog", "rosi", "insu", "chge", "diab", "disc", "adms", "admt", "mod1", "mod2", "mod3", "y")  
dm\_lm <- lm(data = data, formula = nmed ~ diag + proc + hosp\*\*2 + diab + mglu + agem + gend + insu + chge + chge\*labp)  
  
summary(dm\_lm)

##   
## Call:  
## lm(formula = nmed ~ diag + proc + hosp^2 + diab + mglu + agem +   
## gend + insu + chge + chge \* labp, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.8071 -3.4662 -0.4438 3.0748 25.6480   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.163379 1.793874 1.206 0.227893   
## diag 0.944177 0.048594 19.430 < 2e-16 \*\*\*  
## proc 1.204803 0.074823 16.102 < 2e-16 \*\*\*  
## hosp 0.883147 0.031412 28.115 < 2e-16 \*\*\*  
## diabYes 1.348503 0.249786 5.399 7.09e-08 \*\*\*  
## mgluNorm -0.682515 0.174158 -3.919 9.04e-05 \*\*\*  
## agem20-59 2.991256 1.742579 1.717 0.086132 .   
## agem60-79 3.206122 1.741976 1.841 0.065764 .   
## agem80+ 1.863567 1.746984 1.067 0.286154   
## gendMale -0.537621 0.162106 -3.316 0.000919 \*\*\*  
## insuNo -0.587437 0.316028 -1.859 0.063124 .   
## insuSteady -0.359537 0.290061 -1.240 0.215222   
## insuUp 0.304679 0.319155 0.955 0.339813   
## chgeNo -2.582589 0.358723 -7.199 7.14e-13 \*\*\*  
## labp -0.013267 0.008491 -1.562 0.118263   
## chgeNo:labp 0.035090 0.009623 3.647 0.000269 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.184 on 4182 degrees of freedom  
## Multiple R-squared: 0.4288, Adjusted R-squared: 0.4268   
## F-statistic: 209.3 on 15 and 4182 DF, p-value: < 2.2e-16

plot(dm\_lm)



hist(dm\_lm$residuals)



gvlma(dm\_lm)

##   
## Call:  
## lm(formula = nmed ~ diag + proc + hosp^2 + diab + mglu + agem +   
## gend + insu + chge + chge \* labp, data = data)  
##   
## Coefficients:  
## (Intercept) diag proc hosp diabYes mgluNorm   
## 2.16338 0.94418 1.20480 0.88315 1.34850 -0.68252   
## agem20-59 agem60-79 agem80+ gendMale insuNo insuSteady   
## 2.99126 3.20612 1.86357 -0.53762 -0.58744 -0.35954   
## insuUp chgeNo labp chgeNo:labp   
## 0.30468 -2.58259 -0.01327 0.03509   
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma(x = dm\_lm)   
##   
## Value p-value Decision  
## Global Stat 274.258 0.000e+00 Assumptions NOT satisfied!  
## Skewness 156.122 0.000e+00 Assumptions NOT satisfied!  
## Kurtosis 85.814 0.000e+00 Assumptions NOT satisfied!  
## Link Function 2.805 9.395e-02 Assumptions acceptable.  
## Heteroscedasticity 29.517 5.543e-08 Assumptions NOT satisfied!

Per output of glvma, the skewness and kurtosis assumptions are not satisfied. this is due to response variable is skewed to the right.

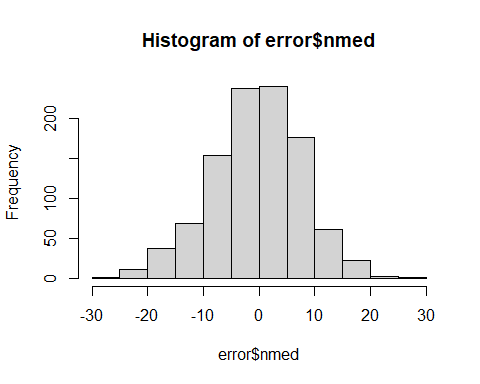
As per Adjusted R-squared the ultimate model only explains 42% of the variation of the number of medications,and p-value: < 2.2e-16 which in my opinion provides too high of residuals to be clinically meaningful.

The model has two quadratic features: age and time in hospital. The time in hospital value is of limited value unfortunately as the data have an artificial maximum of 14 days length of stay. The original analysis omitted hospital lengths of stay >15 days. Additionally, the age versus number of medications has several interesting interpretations. It makes sense that younger people would be on fewer medications, but it appears that the oldest patients are less likely to be on multiple medications. This could be because their doctors are concerned about too many medications or a survivor’s bias.

There were also several dichotomous variables, some of which I adapted. Patients on diabetes medications were much more likely to be on additional medications. By contrast, ‘no changes’ to medications was associated with being on fewer medications, as well as not being insulin-dependent. Men were also own fewer medications.

Finally, I incorporated one interaction into my model. I looked at the number of lab procedures versus medication changes. My rationale for this is that labs would drive the doses of medications to be changed. However, this last interaction does not have a very large impact on the prediction made by the model.

testing$mglu <- testing$mglu %>% recode\_factor('>300' = '>200') #recode testing data  
  
testing <- testing %>% filter(mglu != 'None') %>% drop()  
testing$insu <- testing$insu %>% recode\_factor('Steady' = 'Yes')  
testing$insu <- testing$insu %>% recode\_factor('Up' = 'Yes')  
testing$insu <- testing$insu %>% recode\_factor('Down' = 'Yes')  
  
  
testing$metf <- testing$metf %>% recode\_factor('Steady' = 'Yes')  
testing$metf <- testing$metf %>% recode\_factor('Up' = 'Yes')  
testing$metf <- testing$metf %>% recode\_factor('Down' = 'Yes')  
  
  
prediction <- predict(dm\_lm, data,mu=3)  
  
error <- prediction - testing['nmed']  
  
hist(error$nmed)



t.test(error)

##   
## One Sample t-test  
##   
## data: error  
## t = -1.3665, df = 1010, p-value = 0.1721  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.8421076 0.1507129  
## sample estimates:  
## mean of x   
## -0.3456973

I have used function t.test() that performs T-tests and to calculate T confidence intervals for means.